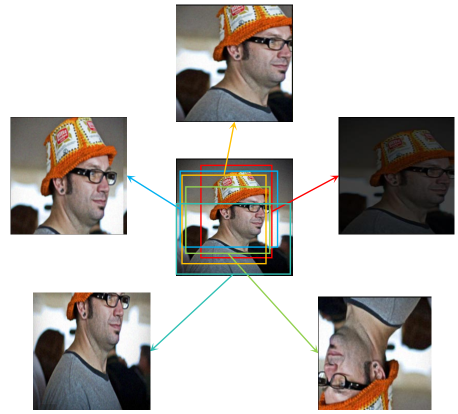
Towards Highly Transferable Vision-Language Attack via Semantic-Augmented Dynamic Contrastive Interaction

Supplementary Material

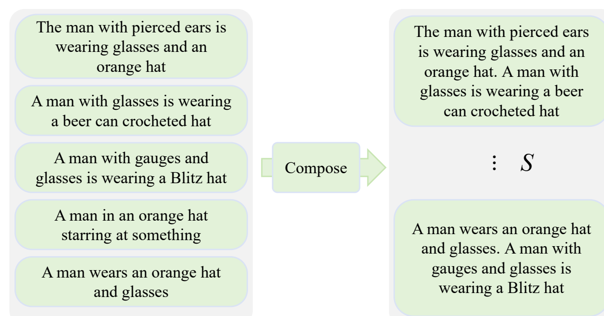
In this paper, we provide an illustration of the semantic augmentation module, experimental results on the MSCOCO dataset, , an evaluation of the effectiveness of the semantic augmentation module, and additional visualizations of the adversarial examples. Furthermore, we also report the attack success rate at Rank-1 (R@1), Rank-5 (R@5), and Rank-10 (R@10) on the Flickr30K dataset.

7. Semantic Augmentation Module

Figure 5 illustrates the Semantic Augmentation Module. For image local semantic augmentation, the module obtains enlarged local regions by randomly cropping and resizing parts of the image, followed by the application of random image augmentation functions to diversify the local semantic content. For text mixed semantic augmentation, the module randomly selects and concatenates pairs of text samples from the text pool to generate a new augmented text set. This approach combines multiple textual descriptions to form broader semantic representations, thereby enhancing semantic diversity.



(a) Local semantic image augmentation



(b) Mixed semantic text augmentation

Figure 5. Semantic Augmentation Module.

8. Experimental Results on MSCOCO dataset

We conduct comparative experiments on the ITR task using four widely adopted VLP models: ALBEF, TCL, CLIP_{ViT} and CLIP_{CNN}. Specifically, each VLP model is used as a source model to generate multimodal adversarial examples, which are then evaluated on the remaining three models to assess cross-model transferability. Table 5 presents the comparative results on the MSCOCO dataset. As shown, SADCA consistently achieves the highest average black-box ASR in both TR and IR tasks. These results demonstrate the effectiveness of SADCA in significantly enhancing the cross-model transferability of multimodal adversarial examples.

9. Ablation Study for Semantic Augmentation Module

To validate the advantages of the Semantic Augmentation Module, we compare it with other input transformation methods. Specifically, we retain the Dynamic Contrastive Interaction (DCI) component and replace the Semantic Augmentation Module with three alternative input transformation methods: DIM [33], SIA [30], and BSR [27]. As shown in Table 6, SADCA equipped with the Semantic Augmentation Module consistently outperforms the alternatives in most cases. This demonstrates that the Semantic Augmentation Module is well-suited to VLP models, effectively enhancing the semantic diversity of inputs and thereby disrupting the alignment mechanisms within vision-language models more efficiently.

10. A Comparison of Attack Cost

Table 7 summarizes the GPU memory usage, runtime, and attack performance when generating adversarial examples using ALBEF as the surrogate model. SADCA achieves the strongest attack performance, reaching 88.35% (TR R@1) and 88.92% (IR R@1), outperforming all compared methods by a clear margin. Although its GPU memory consumption (13.3 GB) and runtime (4.40 h) are moderately higher than those of simpler baselines, they remain substantially lower than the most computationally expensive method, SA-AET(LI)+SIA, while delivering significantly

Table 5. Comparison with SOTA methods on the image-text retrieval (ITR) task on the MSCOCO dataset. The "Source" column indicates the VLP model used to generate the multimodal adversarial examples. For both image retrieval (IR) and text retrieval (TR), we report the ASR (%) at Rank-1 (R@1). The "Average" represents the average ASR on the black-box VLP models.

Source	Attack	ALBEF		TCL		CLIP_{ViT}		CLIP_{CNN}		Average	
		TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
ALBEF	PGD	94.35	93.26	34.15	36.86	21.71	27.06	23.83	30.96	26.56	31.63
	BERT-Attack	24.39	36.13	24.34	33.39	44.94	52.28	47.73	54.75	39.00	46.81
	Co-Attack	94.95	97.87	65.22	72.41	55.28	62.33	56.68	66.45	59.06	67.06
	SGA	99.95	99.94	87.46	88.17	63.72	69.71	63.91	70.78	71.70	76.22
	SGA(LI)+SIA	100	100	98.49	98.01	75.96	79.83	76.75	81.21	83.73	86.35
	DRA	99.90	99.93	88.81	90.06	69.25	75.31	68.53	75.09	75.53	80.15
	SA-AET	100	99.99	97.28	96.88	76.57	80.24	76.17	80.64	85.92	83.34
	SA-AET(LI)+SIA	100	100	**99.66**	**99.50**	91.87	92.69	90.56	92.84	94.03	95.01
	SADCA (ours)	**100**	**100**	98.35	97.99	**93.10**	**94.01**	**93.44**	**95.01**	**94.96**	**95.67**
TCL	PGD	40.81	44.09	98.54	98.20	21.79	26.92	24.97	32.17	29.19	34.39
	BERT-Attack	35.32	45.92	38.54	48.48	51.09	58.80	52.23	61.26	46.21	55.33
	Co-Attack	49.84	60.36	91.68	95.48	32.64	42.69	32.06	47.82	38.85	50.29
	SGA	92.70	92.99	100	100	59.79	65.31	60.52	67.34	71.00	75.21
	SGA(LI)+SIA	99.94	99.43	100	100	74.55	78.94	78.79	82.62	84.43	86.99
	DRA	94.72	95.89	100	100	70.51	74.95	70.29	76.99	78.51	82.61
	SA-AET	97.78	98.08	100	99.99	76.12	79.74	75.89	80.92	83.26	86.25
	SA-AET(LI)+SIA	**99.85**	**99.78**	100	100	90.92	92.92	92.77	94.49	94.51	95.73
	SADCA (ours)	99.59	99.46	**100**	**100**	**93.52**	**94.39**	**95.59**	**96.62**	**96.23**	**96.82**
CLIP_{ViT}	PGD	10.26	13.69	12.72	15.81	82.91	90.51	21.62	28.78	14.87	19.43
	BERT-Attack	20.34	29.74	21.08	29.61	45.06	51.68	44.54	53.72	28.65	37.69
	Co-Attack	26.35	36.69	28.23	38.42	88.78	96.72	47.36	58.45	33.98	44.52
	SGA	43.75	51.08	44.05	51.02	100	100	70.66	75.58	52.82	59.23
	SGA(LI)+SIA	62.49	65.52	64.07	65.21	100	100	93.09	94.80	73.22	75.18
	DRA	52.69	61.50	51.88	61.06	100	100	80.18	84.11	61.58	68.89
	SA-AET	57.64	66.88	57.30	65.16	100	100	83.98	86.72	66.31	72.92
	SA-AET(LI)+SIA	86.23	87.47	85.42	86.16	100	100	97.59	97.96	89.75	90.53
	SADCA (ours)	**90.79**	**91.09**	**88.46**	**87.90**	**100**	**100**	**99.51**	**99.53**	**92.92**	**92.84**
CLIP_{CNN}	PGD	8.38	12.73	11.90	15.68	13.66	20.62	92.68	94.71	11.31	16.34
	BERT-Attack	23.38	34.64	24.58	29.61	51.28	57.49	54.43	62.17	33.08	40.58
	Co-Attack	29.49	41.50	31.83	43.44	53.15	60.15	97.79	98.54	38.16	48.36
	SGA	36.94	46.79	38.81	48.90	62.19	67.70	97.79	98.54	45.98	54.46
	SGA(LI)+SIA	37.15	45.49	39.81	48.26	65.32	72.56	100	100	47.43	55.44
	DRA	41.40	52.25	43.62	54.15	70.43	74.14	99.80	99.92	51.82	60.18
	SA-AET	43.62	55.19	47.01	57.39	73.67	76.90	100	99.92	54.77	63.16
	SA-AET(LI)+SIA	55.35	63.75	58.49	65.66	85.5	87.41	100	100	66.45	72.27
	SADCA (ours)	**62.90**	**69.64**	**63.78**	**69.48**	**88.45**	**91.16**	**100**	**100**	**71.71**	**76.76**

better results. This demonstrates that SADCA offers the most favorable cost–performance balance, providing superior adversarial transferability with a reasonable computational overhead.

11. More Results on Flickr30K Dataset

Reporting only the ASR at R@1 (i.e., the correct image no longer appearing at the top rank) is insufficient for a comprehensive evaluation of the robustness of multimodal retrieval models, as the correct sample may merely be shifted to Rank-2 or Rank-3, which has limited impact on the overall system. In contrast, presenting the attack success rates at R@5 and R@10 provides insight into whether the attack truly disrupts a broader portion of the ranking structure, thereby offering a more complete assessment of the model's vulnerability at the ranking level. Accordingly, we report the ASR at Rank-1, Rank-5, and Rank-10 on the Flickr30K dataset in Table 8 and Table 9. The results show that SADCA achieves strong attack performance across all metrics, indicating its effectiveness in substantially perturbing the retrieval ranking system and demonstrating its stronger

Table 6. Ablation Study for Semantic Augmentation Module.

Source	Attack	ALBEF		TCL		$CLIP_{ViT}$		$CLIP_{CNN}$	
		TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
ALBEF	DCI+DIM	100	100	94.73	94.05	73.37	77.93	74.46	77.77
	DCI+SIA	100	100	**99.58**	**99.57**	70.55	74.58	73.18	78.66
	DCI+BSR	100	100	98.74	98.86	77.30	79.19	78.80	82.20
	SADCA	**100**	**100**	98.52	97.83	**81.10**	**82.83**	**85.44**	**86.11**
TCL	DCI+DIM	97.60	97.26	100	100	75.71	80.54	80.08	83.53
	DCI+SIA	**100**	**99.95**	100	100	71.66	78.38	79.44	83.40
	DCI+BSR	100	99.86	100	100	**79.97**	82.89	86.39	88.30
	SADCA	99.58	99.56	**100**	**100**	78.28	**83.18**	**86.46**	**88.71**
$CLIP_{ViT}$	DCI+DIM	49.95	57.60	48.37	57.57	100	100	80.59	82.64
	DCI+SIA	77.69	80.24	79.45	81.98	100	100	96.93	96.64
	DCI+BSR	79.35	81.55	80.30	82.83	100	100	96.96	97.32
	SADCA	**87.07**	**89.20**	**87.04**	**87.98**	**100**	**100**	**97.90**	**97.46**
$CLIP_{CNN}$	DCI+DIM	23.77	40.34	29.5	43.33	56.69	65.01	100	100
	DCI+SIA	42.46	55.96	47.63	60.02	76.56	80.80	100	100
	DCI+BSR	37.02	51.50	44.26	55.79	69.20	75.58	100	100
	SADCA	**49.43**	**60.55**	**55.53**	**63.19**	**77.18**	**79.57**	**100**	**100**

Table 7. Comparison of attack costs. GPU memory usage, runtime, and attack performance when generating adversarial examples on the Flickr30K dataset using ALBEF as the surrogate model.

Methods	GPU Memory (GB)	Run Time (h)	TR R@1	IR R@1
SGA	10.5	0.83	54.72	61.54
SGA(LI)+SIA	13.0	3.95	67.99	72.07
DRA	10.6	1.57	62.46	69.06
SA-AET	10.5	2.12	69.74	75.17
SA-AET(LI)+SIA	13.5	11.08	83.85	86.12
SADCA	13.3	4.40	88.35	88.92

attack capability.

12. Visualization

Figure 6 shows the visualization of randomly selected clean examples and adversarial examples. Figure 7 shows the description of commercial LVLMs for the adversarial images generated by SADCA. It can be seen that it is capable of effectively attacking various commercial LVLMs.

Table 8. Comparison with SOTA methods on the image-text retrieval (ITR) task on the Flickr30K dataset. The ”Source” column indicates the VLP model used to generate the multimodal adversarial examples. For text retrieval (TR), we report the ASR (%) at Rank-1 (R@1), Rank-5 (R@5) and Rank-10 (R@10).

Source	Attack	ALBEF			TCL			CLIP_{ViT}			CLIP_{CNN}		
		TR R@1	TR R@5	TR R@10	TR R@1	TR R@5	TR R@10	TR R@1	TR R@5	TR R@10	TR R@1	TR R@5	TR R@10
ALBEF	SGA	99.90	99.70	99.70	87.88	77.79	71.74	36.69	19.83	12.40	39.59	**21.88**	14.83
	SGA+SIA	99.79	99.50	99.20	94.10	88.74	85.07	47.48	29.60	21.75	53.77	34.46	24.82
	SGA(LI)	100	100	100	84.19	73.17	67.43	4.36	15.47	10.06	36.53	19.77	12.36
	SGA(LI)+SIA	100	100	100	99.37	98.29	97.80	49.82	32.40	23.68	54.79	35.31	28.01
	DRA	99.90	99.70	99.70	91.57	81.31	75.95	46.26	25.44	18.60	49.55	29.49	21.22
	SA-AET	99.90	99.80	99.80	96.42	92.36	89.98	55.58	34.48	26.32	57.22	39.64	28.53
	SA-AET+SIA	99.58	99.30	98.80	95.21	90.01	85.97	64.54	41.64	34.04	66.16	44.82	36.35
	SA-AET(LI)	100	100	100	98.63	97.29	96.09	60.61	38.11	30.18	62.20	41.12	32.65
	SA-AET(LI)+SIA	100	100	100	**99.58**	**98.89**	**98.50**	75.71	60.12	52.13	76.25	61.21	53.66
	SADCA	**100**	**100**	**100**	98.52	96.28	94.79	**81.10**	**62.10**	**55.08**	**85.44**	**72.09**	**65.91**
TCL	SGA	92.40	87.07	85.40	100	100	99.90	36.81	18.59	13.11	41.89	22.73	14.93
	SGA+SIA	97.18	94.19	91.90	100	99.70	99.60	53.01	33.13	27.03	58.49	41.33	32.34
	SGA(LI)	84.46	76.25	71.50	100	100	100	31.17	15.06	9.86	36.78	20.08	14.01
	SGA(LI)+SIA	99.90	99.70	99.60	100	99.90	99.90	52.87	33.85	26.22	62.45	44.19	36.87
	DRA	95.20	91.28	88.00	100	100	99.80	47.24	26.48	18.90	52.23	30.76	22.97
	SA-AET	98.85	96.79	95.50	100	100	100	56.20	34.68	25.91	59.77	39.22	30.07
	SA-AET+SIA	97.08	93.79	92.60	99.79	99.60	99.20	67.36	49.53	40.65	70.11	54.65	46.14
	SA-AET(LI)	99.37	98.60	98.20	100	100	100	56.20	37.49	28.05	62.32	42.81	32.75
	SA-AET(LI)+SIA	**99.95**	**99.90**	**99.80**	100	100	100	77.04	65.52	57.52	80.20	67.86	61.59
	SADCA	99.58	99.20	99.00	**100**	**100**	**100**	**78.28**	**65.94**	**58.64**	**86.46**	**76.74**	**69.93**
CLIP_{ViT}	SGA	22.42	9.02	5.20	25.08	9.55	6.01	100	100	99.90	53.26	33.83	25.33
	SGA+SIA	44.94	25.65	19.60	47.10	27.24	20.44	99.88	99.48	99.09	81.48	65.12	55.10
	SGA(LI)	21.17	8.92	5.60	23.71	8.74	5.51	100	100	100	49.43	31.40	21.52
	SGA(LI)+SIA	54.33	34.37	27.90	57.43	36.88	30.76	100	100	100	89.02	80.66	72.61
	DRA	27.84	12.73	8.10	27.82	12.46	7.52	100	100	100	64.88	42.49	33.47
	SA-AET	36.60	20.64	16.00	39.20	20.30	14.33	100	100	100	71.01	50.11	41.50
	SA-AET+SIA	54.54	36.17	30.20	58.59	38.79	31.96	99.63	99.48	98.98	83.91	69.34	62.00
	SA-AET(LI)	45.57	28.76	21.80	46.89	27.44	21.14	100	100	100	78.67	60.89	52.01
	SA-AET(LI)+SIA	79.04	64.33	58.40	79.35	66.33	59.72	100	100	100	94.76	89.32	85.79
	SADCA	**87.07**	**74.85**	**68.40**	**87.04**	**74.07**	**66.93**	**100**	**100**	**100**	**97.90**	**92.81**	**89.91**
CLIP_{CNN}	SGA	15.64	5.61	3.00	18.02	6.03	2.91	39.92	20.15	13.52	99.87	99.47	99.07
	SGA+SIA	19.29	6.11	4.00	22.76	8.34	5.21	46.50	23.88	16.16	99.62	97.99	96.81
	SGA(LI)	15.51	5.61	2.50	17.70	6.13	3.51	38.16	18.38	12.30	99.87	99.89	99.90
	SGA(LI)+SIA	20.96	9.02	5.40	22.55	11.06	7.21	47.24	26.27	19.41	100	99.68	99.38
	DRA	19.50	6.31	3.70	21.60	7.54	3.81	48.47	99.47	17.89	99.87	25.13	99.38
	SA-AET	23.98	9.22	6.00	27.29	10.45	6.61	54.11	33.33	24.39	100	100	99.90
	SA-AET+SIA	26.80	12.63	8.70	31.61	13.77	8.52	57.18	35.62	25.20	99.49	98.31	96.70
	SA-AET(LI)	31.07	15.03	12.00	33.19	15.08	11.62	61.72	43.09	34.45	100	100	100
	SA-AET(LI)+SIA	38.69	20.74	17.10	44.89	25.13	19.74	69.57	52.34	42.89	100	99.79	99.79
	SADCA	**49.43**	**26.05**	**20.90**	**55.53**	**30.15**	**23.45**	**77.18**	**56.39**	**48.48**	**100**	**100**	**100**

Table 9. Comparison with SOTA methods on the image-text retrieval (ITR) task on the Flickr30K dataset. The "Source" column indicates the VLP model used to generate the multimodal adversarial examples. For image retrieval (IR), we report the ASR (%) at Rank-1 (R@1), Rank-5 (R@5) and Rank-10 (R@10).

Source	Attack	ALBEF			TCL			CLIP_{ViT}			CLIP_{CNN}		
		IR R@1	IR R@5	IR R@10	IR R@1	IR R@5	IR R@10	IR R@1	IR R@5	IR R@10	IR R@1	IR R@5	IR R@10
ALBEF	SGA	99.93	99.92	99.90	88.05	77.65	71.32	46.78	29.29	22.18	49.78	32.70	24.80
	SGA+SIA	99.79	99.41	99.15	94.21	88.11	83.33	57.38	38.68	30.88	61.23	43.96	35.19
	SGA(LI)	100	100	100	84.50	72.94	66.21	41.04	23.41	17.60	45.18	26.95	20.71
	SGA(LI)+SIA	100	100	100	99.02	97.40	96.28	56.19	38.03	30.77	60.99	43.86	34.63
	DRA	99.93	99.86	99.82	91.36	82.24	77.03	56.80	38.61	29.79	59.01	41.65	33.52
	SA-AET	99.98	100	99.98	96.02	92.02	89.05	63.89	45.88	36.52	65.59	48.84	40.03
	SA-AET+SIA	99.72	99.36	98.99	95.21	90.01	85.97	70.04	53.47	45.12	72.25	56.55	47.73
	SA-AET(LI)	100	100	100	98.33	96.63	95.25	66.91	49.22	40.80	69.09	51.92	43.14
	SA-AET(LI)+SIA	100	**100**	99.98	**99.38**	**98.78**	**98.35**	78.58	66.36	59.70	80.41	68.15	60.49
	SADCA	**100**	99.98	**99.98**	97.83	94.60	92.65	**82.83**	**68.89**	**61.49**	**86.11**	**75.30**	**68.92**
TCL	SGA	92.77	87.08	83.99	100	99.96	99.96	46.97	28.97	22.15	51.53	33.26	25.55
	SGA+SIA	97.22	94.13	92.38	99.98	99.75	99.65	61.76	44.52	36.09	65.21	49.76	41.72
	SGA(LI)	85.24	75.72	70.91	100	100	100	38.92	23.27	17.23	45.97	27.49	20.92
	SGA(LI)+SIA	99.70	99.28	99.03	100	99.98	99.98	59.12	42.37	35.41	65.49	50.75	42.58
	DRA	95.58	91.10	88.66	100	99.94	99.94	57.28	39.90	32.16	62.23	43.55	35.19
	SA-AET	98.50	96.60	95.45	100	100	99.96	63.47	46.55	38.64	67.86	49.90	41.52
	SA-AET+SIA	97.36	94.48	92.28	99.86	99.38	99.21	72.97	57.65	50.52	75.40	61.77	53.94
	SA-AET(LI)	99.20	98.61	98.08	100	100	100	64.37	47.44	40.10	69.30	51.94	44.18
	SA-AET(LI)+SIA	**99.93**	**99.82**	**99.66**	100	100	100	81.48	69.75	63.65	84.05	73.85	67.47
	SADCA	99.56	98.79	98.28	**100**	**100**	**100**	**83.17**	**70.61**	**64.17**	**88.71**	**79.69**	**74.38**
CLIP_{ViT}	SGA	34.59	18.27	13.99	36.45	19.43	14.11	100	100	100	61.10	43.50	35.83
	SGA+SIA	55.22	35.42	28.22	56.48	37.44	29.89	99.77	99.53	99.24	83.12	69.97	62.34
	SGA(LI)	31.08	15.34	11.04	33.52	16.85	11.76	100	100	100	56.23	37.63	30.63
	SGA(LI)+SIA	60.20	41.82	33.64	62.60	43.64	36.45	100	100	100	89.24	81.00	74.95
	DRA	42.84	25.47	19.29	44.60	25.93	19.60	100	100	99.93	69.50	54.46	46.15
	SA-AET	50.44	32.98	26.14	51.10	33.16	25.87	100	99.93	99.93	74.10	60.24	52.72
	SA-AET+SIA	65.13	47.27	40.80	66.33	49.14	42.06	99.87	99.32	98.95	85.69	74.82	68.74
	SA-AET(LI)	57.16	39.60	32.57	57.52	39.47	31.90	100	100	100	80.17	66.50	59.59
	SA-AET(LI)+SIA	82.74	70.14	63.89	82.57	70.73	64.38	99.97	99.84	99.83	95.23	90.83	87.51
	SADCA	**89.20**	**75.84**	**68.90**	**87.98**	**73.91**	**67.00**	**100**	**99.98**	**99.98**	**97.46**	**93.98**	**91.48**
CLIP_{CNN}	SGA	28.06	15.01	10.68	33.07	16.66	12.02	51.45	32.19	25.27	99.90	99.73	99.64
	SGA+SIA	33.63	17.49	12.66	37.57	20.67	15.21	54.41	36.72	38.96	99.73	98.93	97.90
	SGA(LI)	28.02	14.03	9.97	31.36	15.77	11.19	48.90	29.57	22.57	99.93	99.90	99.80
	SGA(LI)+SIA	32.72	17.19	12.52	37.21	20.55	15.11	56.70	37.86	30.92	100	99.95	99.86
	DRA	34.59	18.56	13.93	37.88	21.15	15.98	59.12	99.64	32.71	99.90	40.32	99.41
	SA-AET	38.28	21.47	16.60	41.81	24.62	18.03	64.21	44.34	36.46	99.97	99.71	99.50
	SA-AET+SIA	41.53	24.65	19.07	46.62	28.24	21.77	64.11	46.90	38.70	99.18	97.96	97.18
	SA-AET(LI)	44.74	27.42	21.71	47.45	29.48	23.09	68.52	52.32	44.29	100	100	100.00
	SA-AET(LI)+SIA	51.80	33.65	26.95	56.33	38.00	31.07	74.68	59.87	52.33	100	99.93	99.89
	SADCA	**60.55**	**38.31**	**30.89**	**63.19**	**41.29**	**33.10**	**79.57**	**64.07**	**56.17**	**100**	**100**	**100**

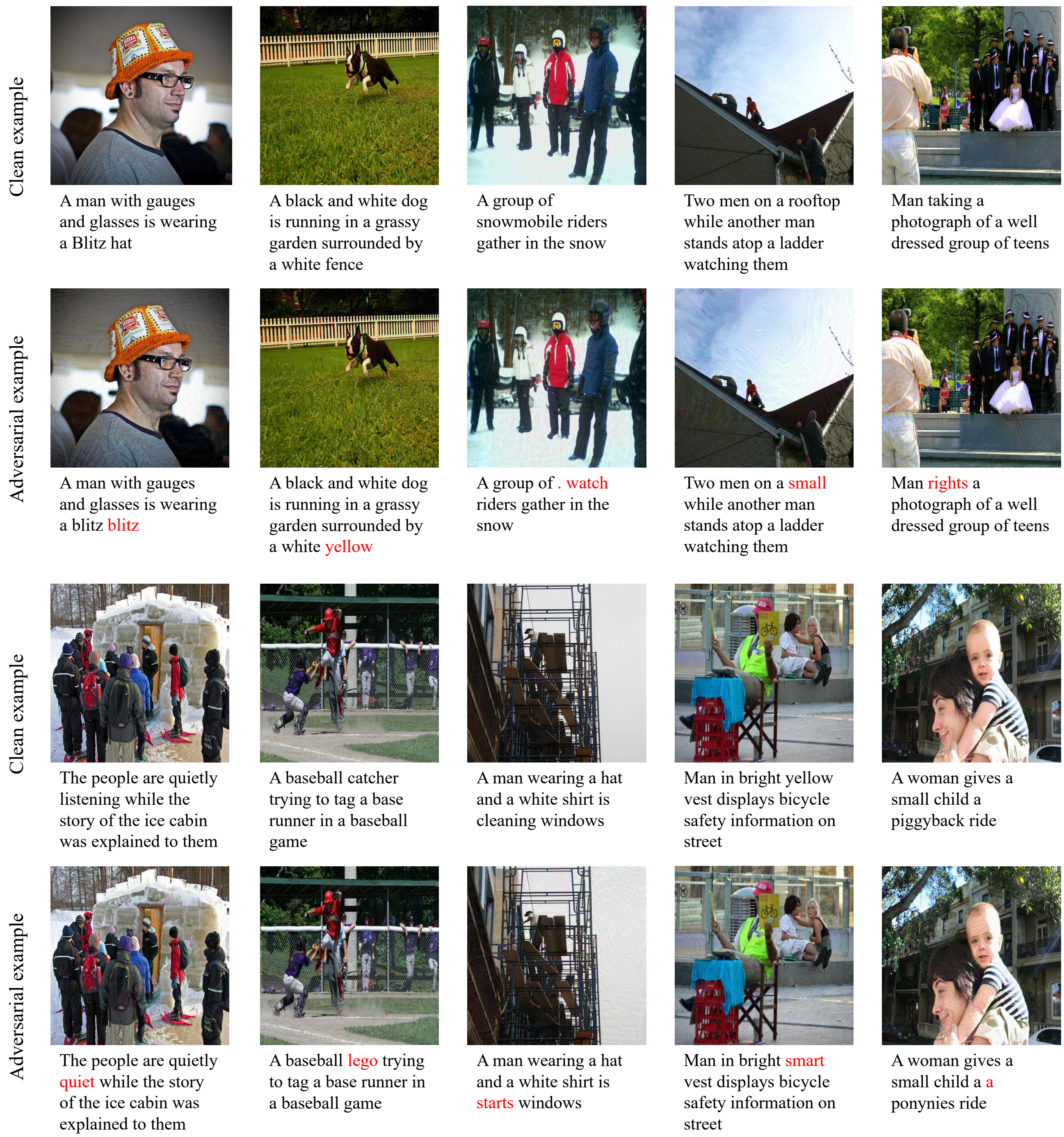


Figure 6. Visualization of original images and the corresponding adversarial examples generated by our proposed SADCA.

Please describe this image, limited to 50 words.

Adversarial Image

Group stands around a table outdoors at night, wearing long-sleeve shirts and jackets. They hold white boxing gloves, appearing to playfully spar. The table has brown leaf-shaped items on a checkered surface. Background shows dark setting with scattered objects and bottles. Casual, lively moment.

GPT-5

Three people gather around an outdoor grill at night, cooking meat. They hold bowls and appear to be enjoying a casual barbecue. The scene shows a relaxed, social atmosphere with friends or family sharing a meal together in a dimly lit outdoor setting.

Claude-4.5

Three people are gathered around what appears to be a grill. A woman holds a plate, a child holds a bowl, and another person is leaning over the grill. The scene is dimly lit, with a table of bottles visible in the background.

Gemini-2.5

Please describe this image, limited to 50 words.

Adversarial Image

An outdoor market or booth scene with a person in a dark zippered jacket sitting at a table displaying items. There are gloves, a purse, a stuffed toy or knitted object, spray bottles, paper towels, and a chalkboard sign in the background. Sunlit urban setting with structures and greenery.

GPT-5

A woman in a dark top stands in a kitchen, smiling at the camera. Behind her is a chalkboard with "Friends to Make" written on it. The kitchen counter shows cooking activity with pots, utensils, and ingredients. A casual, homey cooking scene.

Claude-4.5

A young man with light brown hair and a green hoodie smiles at the camera, sitting at a table. Behind him, a chalkboard reads "MISONA ITS MATE." Bottles and glasses are on the table, with an outdoor area featuring trees and structures visible in the background. The image has a slightly textured, almost painterly quality.

Gemini-2.5

Please describe this image, limited to 50 words.

Adversarial Image

A child in a dark blue hood sits or is held by an adult. A white textured blanket or costume fills foreground. Background shows colorful indoor or outdoor setting with a yellow pillar, decorations, and other people. Overall warm, busy scene, possibly at a parade, festival, or fair.

GPT-5

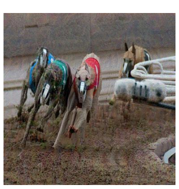
A man with traditional Māori tā moko (facial tattoos) looks directly at the camera while surrounded by others. The image appears to capture a cultural ceremony or gathering, with white fabric visible in the foreground and people in the background against a yellow wall.

Claude-4.5

A man with short blonde hair and a plaid shirt looks towards the viewer with a slightly surprised expression. He wears a tie. The background is blurred, showing hints of other people and a yellow surface. The entire image has a distinct, swirling, dream-like filter applied, giving it an ethereal and somewhat distorted quality.

Gemini-2.5

Please describe this image, limited to 50 words.



Adversarial Image

Grainy, surreal photo with dreamlike artifacts. A hooded person bends forward in the center, holding a dark object. To the left, two bicycles or scooters lean against a wall. On the right lies a large coil of white rope or hose and cardboard. Earthy ground and muted urban background.

GPT-5

A person in tan/beige clothing stands in a dusty outdoor area, holding what appears to be rodeo equipment including a pink/red item. Behind them are metal gates or fencing, with a horse visible in the background. The setting suggests a rodeo or ranch environment.

Claude-4.5

This abstract, distorted image features blurred figures and organic textures. On the left, blue and green shapes suggest a backpack or person. A central figure wears light clothing with a red element. The background shows a wall, while the foreground is an earthy, textured surface with hints of foliage, creating a surreal scene.

Gemini-2.5

Figure 7. Visualization of adversarial images in attacking commercial LVLMs.